Harmonic Coding: A Low Bit-Rate, Good-Quality Speech Coding Technique

by

Luis B. Almeida

and

José M. Tribolet

Centro de Análise e Processamento de Sinais
Complexo I, 1096 LISBOA CODEX, PORTUGAL
Tel. 572399

ABSTRACT

Low bit-rate, good-quality speech coding is one of the fundamental goals of today's speech processing research. Present-day coding techniques, like APC and ATC, are able to achieve good-quality transmission only down to about 12 kb/s. Below this rate, their quality degrades rapidly. On the other hand, the various kinds of vocoders, which operate up to about 5 kb/s, have inherent quality limitations which cannot be overcome by an increase of the bit rate.

In this paper, a new coding scheme is presented, which is based on a recently developed spectral model for nonstationary voiced speech, and it forms the basis of a waveform coder and a vocoder which are introduced in this paper, and which share the same basic structure.

Experimental results are presented, which show that both systems yield significant bit-rate reductions relative to present-day schemes of equivalent quality.

I. Introduction

Low bit-rate, good-quality speech coding is one of the fundamental goals of today's speech processing research. The design of toll-quality speech coders for commercial data-rate operation seems particularly important at this point in time, to allow extended services on existing telephone networks.

Digital transmission of speech can be achieved essentially by two different methods. In waveform coding methods [1,2,3], the emphasis lies on the reproduction of the speech waveform itself, according to some fidelity criterion. In vocoding methods [4] the emphasis lies on the reproduction of the speech production mechanism.

Either scheme has important advantages and disadvantages. Waveform coding schemes are inherently capable of toll-quality, while vocoding schemes are not. Vocoders are however capable of reasonable communications quality at much lower transmission rates than waveform coders.

The present state-of-the-art is capable of yielding toll-quality waveform coded speech at bit rates of 12 kb/s or higher. Below this rate, however, waveform coders are known to degrade progressively, yielding unacceptable performance at rates of 4.8 kb/s and below, at which rates the best vocoding schemes yield vastly superior results.

For several years now, there have been attempts to close the performance gap between vocoders and waveform coders with various degrees of success. In any case, the goal of toll-quality transmission in the range from 4.8 kb/s to 9.6 kb/s still remains as elusive as ever.

In this paper, a novel coding scheme, called *Harmonic Coding* is presented. The Harmonic Coding philosophy is a symbiose of waveform coding and vocoding philosophies, and comes as a direct result of a novel non-stationary spectral model for voiced speech recently developed by the authors [5].

This paper is organized as follows: Section II presents a brief overview of the non-stationary spectral model for voiced speech. In Section III the general Harmonic Coding scheme is described, while in Section IV a particular 4.8 kb/s architecture is presented and its operation illustrated. Experimental results will then be discussed in Section V.

II. Harmonic Model

A novel nonstationary spectral model for voiced speech has been recently introduced [5], according to which the short-time spectrum of voiced speech $S(t,\omega)$ can be represented as a sum of generalized harmonics centered at multiples of the instantaneous fundamental frequency $\Omega(t)$, as:

$$S(t,w) = \sum_{r=-\infty}^{\infty} \sum_{\ell=0}^{\infty} K_{r\ell}(t) A^{*(\ell)}[\omega - r\,\Omega(t)] \qquad (1)$$

where:

- $S(t,\omega)$ is the Sliding-Origin short-time Fourier transform of voiced speech $s(t)$, using an analysis window $a(t)$.
- $\{K_{r\ell}(t), \ell=0,1,...\}$ are the coefficients of the r^{th} generalized harmonic.
- $A^{*(\ell)}$ is the ℓ^{th} derivative of the conjugate Fourier transform of $a(t)$.

The model of Eq. (1) is capable of describing the short-time spectrum of any signal $s(t)$, being it periodic or not.

For periodic or quasi-periodic signals however, $S(t,\omega)$ can be quite accurately represented with a truncated model of a low order L, for which all harmonic coefficients $\{K_{r\ell}(t), \ell > L\}$ in Eq. (1) are set to zero.

Experimental studies reported elsewhere [5] showed that voiced speech is quite well represented by a 2nd order model. In other words, a pure analysis/synthesis scheme based on a 2^{nd} order generalized harmonic model yields good-quality speech. Further studies showed that zero-order models introduced a slight degradation in speech quality, which is however barely noticeable.

Along with the model of Eq. (1), several explicit, non-linear prediction relationships relating the magnitudes and phases of the generalized harmonics at various instants in time were derived [5]. In particular a simple, yet reasonably accurate *phase prediction* relationship was derived for zero-order harmonic models.

CH 1746-7/82/0000 - 1664 $ 00.75 © 1982 IEEE

Due to its inherent data compression capabilities, the zero-th order model was chosen as the basis for the 4.8 kb/s harmonic coding architecture to be described in Section IV.

III. General Description of an Harmonic Coder

Figure 1 depicts the general diagram of an harmonic coder. At the transmitter the data is pre-filtered, framed and windowed and transformed to the frequency domain, to yield short-time spectra $S(n,k)$.

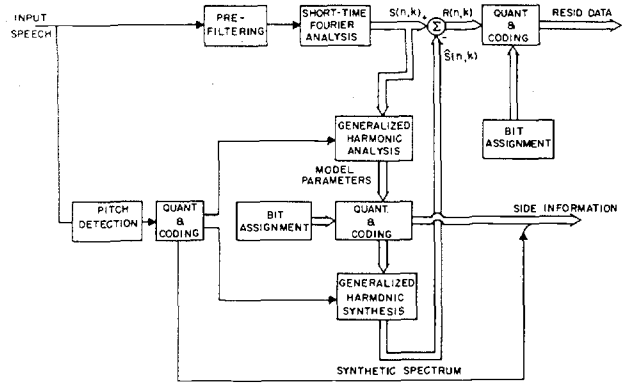


Fig. 1 - General Harmonic Coder diagram.
a - Transmitter. b - Receiver.

The short-time spectrum is then analyzed into generalized harmonics, according to the current estimate of the instantaneous pitch. The model parameters, i.e., the complex amplitudes of the generalized harmonics, are then quantized and used to synthesize the modeled spectrum $\hat{S}(n,k)$, according to Eq. (1).

The residual spectrum $R(n,k) = S(n,k) - \hat{S}(n,k)$ is then quantized and sent to the receiver, along with the pitch and model coefficients, as side information.

At the receiver the residual data are decoded and added to the synthetic spectrum, and then fed to the short-time Fourier synthesis and post-filtering.

The proposed scheme bears a strong resemblance to Adaptive Transform Coding [2,3]. It is however a much upgraded version for two main reasons. The first is concerned with the fact that this scheme uses a better spectral model than previous ATC schemes. The second is that it can be gracefully downgraded to low-bit rate transmission without any bandwidth loss and ultimately operate as a vocoder, as detailed in the next section.

The major improvement in modeling the short-time spectrum is mainly due to the way the phase component is dealt with. It is this fact that allows subtracting the modeled spectrum from the speech spectrum, since they are essentially in phase.

The proposed system is obviously an identity system, in the absence of quantization effects. Quantization introduces two sources of degradation, namely, in the representation of the modeled spectrum and in the representation of the residual spectrum. There are thus two bit assignment strategies to contend with, as illustrated in Figure 1. The choice of a specific strategy is of course dependent on transmission rate.

The model coefficients, which are complex numbers, can be represented in various ways, e.g., in rectangular or polar coordinates. Polar seems a good choice in that it allows the independent handling of amplitude and phase quantization noise, which are known to have subjectively distinct effects.

Many variations of the basic scheme are, of course, possible. Our research has concentrated, so far, on understanding and optimizing the system performance at 4.8 kb/s, as shall be detailed next.

IV. An Architecture for 4.8 kb/s Harmonic Coding

We have studied a 4.8 kb/s Harmonic Coder characterized by the following properties:

1 — A harmonic spectral model of order zero was used.

2 — 12-pole LPC pre-filtering was used as a spectral flattener.

3 — No bits were assigned for transmitting the residual spectrum $R(n,k)$.

resulting in the basic architecture shown in Figure 2.

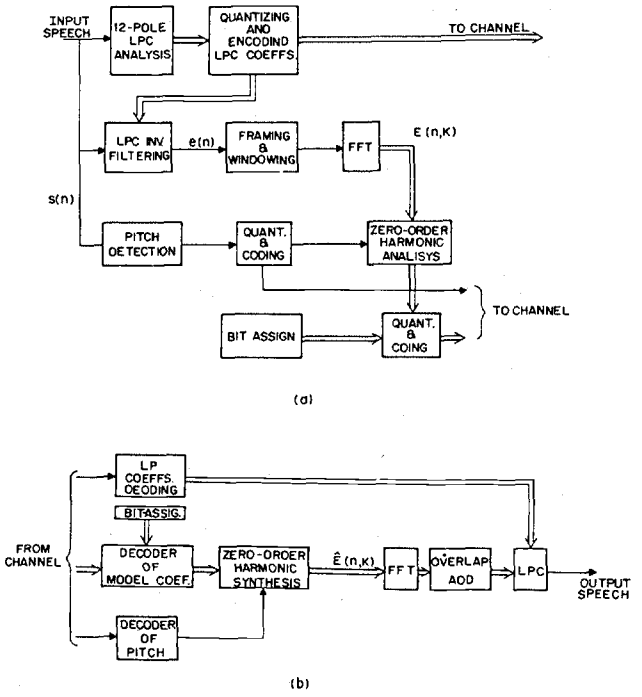


Fig. 2 - Structure of the 4.8 kb/s Harmonic Coder
a - Transmitter b - Receiver

For analysis, we have used a Hanning window, 32 msecs. long, with 50% overlap between frames. The 12 LPC coefficients are assigned 40 bits and are transmitted every other frame. Gain, pitch and mean fundamental frequency (which is used for phase prediction [5,6]) use 18 bits per frame. The model coefficients use then a total of 39 bits per frame.

If the number of bits for model coefficient representation is progressively reduced, the system quality gracefully degrades until, at zero bits/coefficient it becomes essentially equivalent to an LPC vocoder.

So far we have only concentrated on the processing of voiced speech. Unvoiced speech is dealt with very simply, in a way similar to that of an LPC vocoder, except that the random excitation is generated in the frequency domain, rather than in the time domain.

Figure 3 depicts the operation of the coder in a voiced segment. As can be seen from Fig. 3-f, no amplitude information was transmitted: harmonic amplitudes were synthesized as constant at the receiver. Fig. 3-g shows the efficacity of phase prediction, while Figs. 3-h,i show phase transmission. As can be seen from Figs. 3-a,b, input and output speech are in phase.

V. Coder Evaluation

Figure 4 shows sample spectrograms of input and output speech (sentence: "A lathe is a big tool"), for a male speaker. Since this vocoder is a preliminary version of an improved version presently under development, no formal subjective quality tests have been made yet. Informal listening shows that the output speech has good communications quality. Some tonal distortion, harshness and buzziness are present, especially for male speakers. This version of the 4.8 kb/s harmonic coder however provided much insight as to how the model coefficients should be represented and coded in a low bit-rate environment. Further developments are expected to yield significant quality improvements.

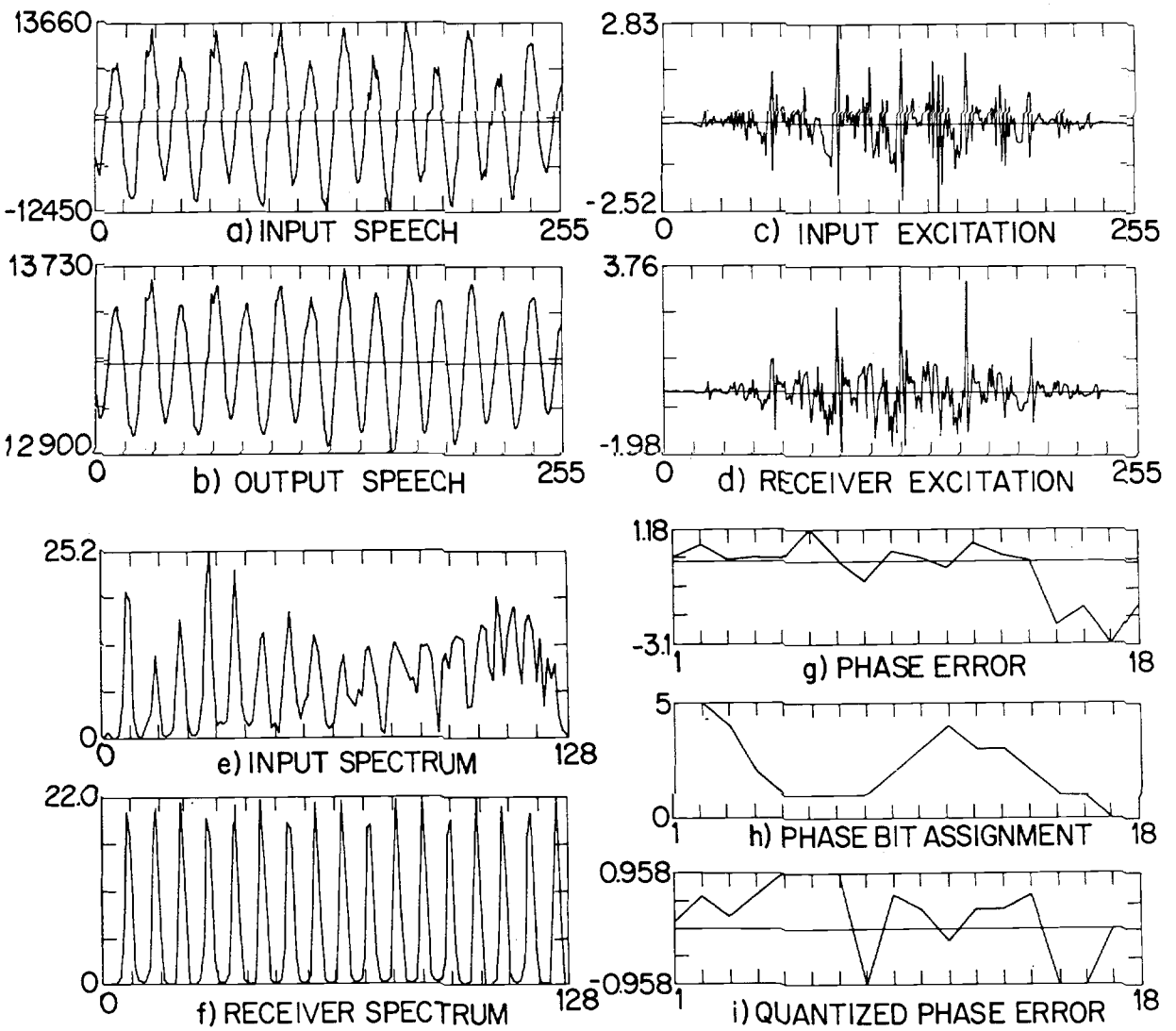
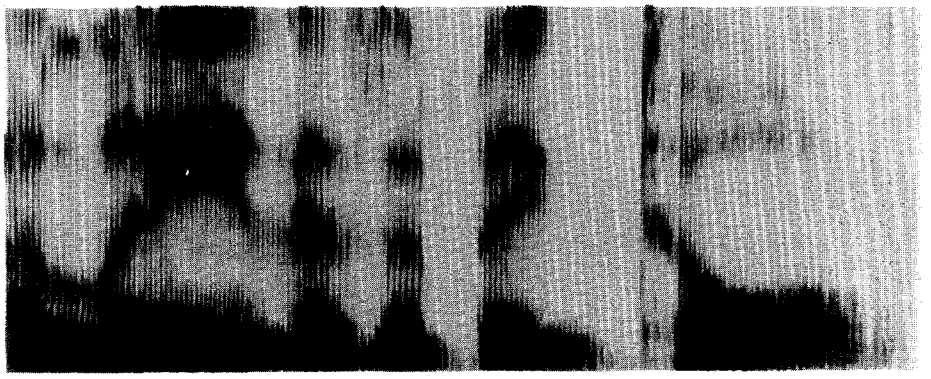
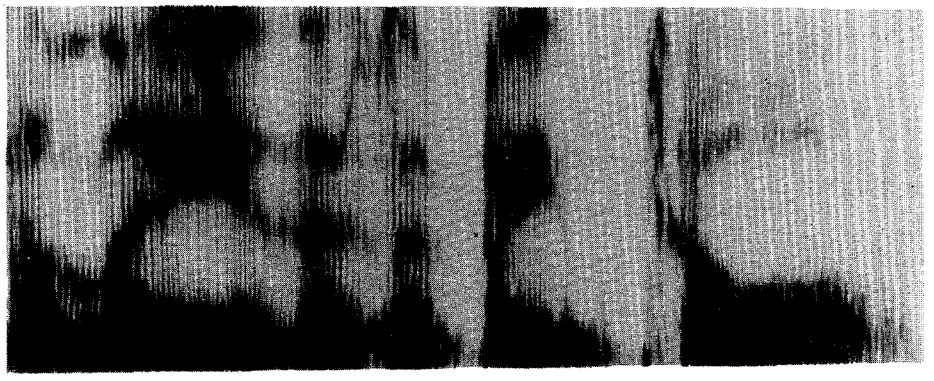


Fig. 3 - Illustration of the operation of the 4.8 kb/s Harmonic Coder



(a) - Original speech



(b) - Output speech

Fig. 4 - Illustration of the performance of the 4.8 kb/s Harmonic Coder

VI. Conclusions

A new, general coding scheme, based on a model which provides an efficient representation of the spectrum of voiced speech, appears very promising in yielding good- to high-quality speech reproduction in the range from 4.8 kb/s to 9.6 kb/s. Systems of this type, being very flexible, easily lend themselves to variable-rate speech coding applications, with graceful degradation as the bit rate is progressively reduced.

This is a preliminary report. Much remains to be done in understanding the operation of this class of coders, one of the main issues being the representation, bit-assignment and quantization of coefficient amplitudes and phases.

References

[1] B. S. Atal and M. R. Schroeder, "Adaptive Predictive Coding of Speech Signals," Bell Syst. Tech. J., Vol. 49, October 1970, pp. 1973-1986.

[2] J. M. Tribolet and R. E. Crochiere, "Frequency Domain Coding of Speech," IEEE Trans. Acoust., Speech, Signal Processing, Vol. ASSP-27, Oct. 1979, pp. 512-530.

[3] R. V. Cox and R. E. Crochiere, "Real-Time Simulation of Adaptive Transform Coding," IEEE Trans. Acoust., Speech, Signal Processing, Vol. ASSP-29, April 1981, pp. 147-154.

[4] J. L. Flanagan, "Speech Analysis, Synthesis and Perception," New York: Springer-Verlag, 1972.

[5] L. B. Almeida and J. M. Tribolet, "A Spectral Model For Nonstationary Voiced Speech," Proc. of the ICASSP 82, Paris, April 1982.

[6] L. B. Almeida and J. M. Tribolet, "A Model for Short-Time Phase Prediction of Speech," Proc. of the ICASSP 81, Atlanta, March 1981.